Journal of Official Statistics
Vol. 5, No. 4, 1989, pp. 365–374
© Statistics Sweden

Probability Estimation With Sample Compositing Techniques

Forest C. Garner,[1] Martin A. Stapanian,[1] Evangelos A. Yfantis,[2] and Llewellyn R. Williams[3]

Abstract: Sometimes the objective of an experiment is to estimate the proportion p of individuals that possess a characteristic, such as a parasite, a blood disease, or an antibody, and it is not necessary to identify those individuals. In such a case, analyzing composites of aliquots from k individual test portions may result in an estimator of p with substantially lower mean square error than the traditional estimator. The mean square error varies with k and p, and we show a method for choosing the optimal value of k. Practical considerations for environmental scientists designing such an experiment are discussed. Substantial cost savings may result when a carefully planned sample compositing experiment is used.

Key words: Sample compositing; group testing; probability estimation.

1. Introduction

Suppose the objective of an experiment is to estimate the proportion of individuals possessing a characteristic, such as a blood type, a disease, a parasite, or an enzyme. Suppose further that it is not necessary to identify the individuals that are positive for the characteristic. Sample compositing, or group testing, could be applied to this experiment. After taking sufficient test material from the appropriate number of individuals or sites, the individual test portions, or aliquots (subportions) from the test portions may be combined. The resulting composites would then be analyzed. In a traditional experiment, each test portion would be analyzed separately. Substantial savings in analytical costs may occur when a carefully planned compositing strategy is used.

General applications of sample compositing include (1) estimation of parameters, particularly the mean; (2) identification of defective, or positive individuals; and (3) estimation of proportions. Specific examples include sampling plankton (Cassie (1971)), bales of wool (Cameron (1951)), determining levels of insecticide in fruit (Ryan, Pilan,

Acknowledgements: We thank K.E. Fitzgerald for assistance in computer programming. C.C. Smith, J.J. Higgins, C. Chen, and two anonymous reviewers provided useful comments on the manuscript. Although the research described in this article has been funded wholly or in part by the U.S. Environmental Protection Agency through contract 68-03-3249 to Lockheed Engineering & Sciences Co., it has not been subject to Agency review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

[1] Lockheed Engineering & Sciences Co., 1050 E. Flamingo Rd., Las Vegas, NV 89119, U.S.A.
[2] Department of Computer Science, University of Nevada, Las Vegas, Las Vegas, NV 89154, U.S.A.
[3] Exposure Assessment Research Division, United States Environmental Protection Agency, Environmental Monitoring Systems Laboratory, 944 E. Harmon Avenue, Las Vegas, NV 89109, U.S.A.

and Leduc (1982)), fat content in milk (Connolly and O'Connor (1982)), and characterizing the communication of disease (Thompson (1962)). Sample compositing has many applications in industry (Sobel and Groll (1959)) and in geochemistry and remote sensing (Garrett and Sinding-Larsen (1984)). Advantages of sample compositing include the reduction of experimental costs (Watson (1961)), reduction in the variance of an estimate of average concentration (Garner, Stapanian, and Williams (1987)), and increase in the precision and probability of detection (Mack and Robinson (1985)). Rohde (1976) attempted to describe the distribution of composite analyses when the test portion sizes are random.

In spite of these advantages and widespread applications, sample compositing is used rather infrequently by most environmental scientists. Environmental studies generally have severe budgetary constraints. As a result, traditional sampling techniques, which generally require less experimental planning, are used without consideration of sample compositing. Sample compositing for estimating probability is not common, and most of the relevant literature is relatively unknown or not readily accessible to most environmental scientists. The purpose of this paper is to direct the attention of environmental scientists to the use of sample compositing for estimating probability. Mathematically this paper is similar to those of Gibbs and Gower (1960), Thompson (1962) and Sobel and Elashoff (1975).

In this paper, we discuss and compare characteristics of traditional experiments and composite experiments. In a typical experiment using sample compositing, the n test portions are partitioned into n/k distinct sets. Equal-sized aliquots are taken from each test portion. The aliquots from each set of k test portions are physically combined to form n/k composites which are each subsequently analyzed. Limitations of such composite experiments are discussed. We describe the relationship between the proportion of the population having the characteristic and (1) its traditional estimator and (2) its maximum-likelihood estimator when sample compositing is used. The mean square errors of the estimators are compared for a fixed number of analyses. The costs of traditional and composite plans are compared. We assume that proper procedures are used in composite experiments, such as ensuring homogeneity of each test portion. The model below assumes that the characteristic of interest is either present in detectable amounts or absent. The results of an analysis are either positive or negative. The effects of "false positive" and "false negative" rates are discussed.

2. The Estimator

Let p be the proportion of individuals possessing a characteristic. Suppose n individuals are chosen randomly from an (essentially) infinite population. A representative test portion is taken from each individual.

Under these conditions, the parameter p can be estimated in the traditional way by separately analyzing the test portion from each individual, and by using

$$\hat{p}_1 = X_1/n \tag{2.1}$$

as an estimator of p, where X_1 is the number of individual test portions that test positive. The mean and variance of the traditional estimator are, respectively,

$$E[\hat{p}_1] = p \tag{2.2}$$

and

$$V[\hat{p}_i] = E[(\hat{p}_1 - p)^2] = p(1-p)n^{-1}. \tag{2.3}$$

The cost of obtaining this estimate of p is

$$C_1 = n(C_s + C_A), \tag{2.4}$$

where C_1 is the total cost, C_s is the cost of obtaining a test portion from one individual, and C_A is the cost of performing one analysis. Because of (2.2), $\hat{p}_1$ is an unbiased estimator of p. The estimator $\hat{p}_1$ has the least variance among all unbiased estimators of p under these conditions.

Suppose random samples of size k are chosen from the n individual test portions ($n \gg k$), and each group of test portions is composited and analyzed. Let $m = n/k$ be the number of composites. Let X_k be the number of composites that test positive. The random variable X_k has a binomial distribution with density

$$f(x) = \binom{m}{x} p'^x (1 - p')^{m-x},$$

$$x = 0, 1, 2, \ldots, m \tag{2.5}$$

where p' is the probability that a composite possesses the characteristic. The relation between p' and p is

$$p' = 1 - (1 - p)^k. \tag{2.6}$$

Thus,

$$p = 1 - (1 - p')^{1/k}. \tag{2.7}$$

The maximum-likelihood estimator of p is

$$\hat{p}_k = 1 - (1 - X_k/m)^{1/k}. \tag{2.8}$$

The mean and variance of $\hat{p}_k$ are, respectively,

$$E(\hat{p}_k) = \sum_{i=0}^{m} [1 - (1 - i/m)^{1/k}] f(i) \text{ and} \tag{2.9}$$

$$V(\hat{p}_k) = \sum_{i=0}^{m} [1 - (1 - i/m)^{1/k}]^2 f(i) - [E(\hat{p}_k)]^2. \tag{2.10}$$

The cost of the composite sampling plan is

$$C_k = n(C_s + C_A/k). \tag{2.11}$$

The reduction in cost is emphasized when C_A is substantially greater than C_s.

We use the mean square error (MSE) as a measure of the quality of $\hat{p}_k$ and $\hat{p}_1$.

$$\text{MSE} = \sigma^2 + (\mu - p)^2 \tag{2.12}$$

where σ^2 and μ are the variance and mean, respectively, of the estimators of p. The MSE is used to compare the traditional (2.1) and the composite (2.8) estimators because the traditional is unbiased and the composite is biased. The costs of traditional and composite experiments should be compared for a fixed MSE.

This theory can be extended to encompass measurement errors. Typically there is a small probability α of incorrectly detecting the characteristic when it is not present (a false positive) and a small probability β of failing to detect the characteristic when it is present (a false negative). Thus, the frequency of detection, p_d, is related to the frequency of occurrence, p', by

$$p_d = (1 - p')\alpha + p'(1 - \beta). \tag{2.13}$$

The performance of most measurement methods in wide usage is well known, and very good estimates of α and β are generally available. The true frequency may be estimated by

$$\hat{p}' = (\hat{p}_d - \alpha)(1 - \alpha - \beta)^{-1} \tag{2.14}$$

where $\hat{p}_d$ is the observed frequency of detection. The probability of occurrence of the characteristic of interest in a single test portion is estimated by

$$\hat{p}_k = 1 - (1 - \hat{p}')^{1/k}. \tag{2.15}$$

For many environmental analytical methods, α and β are small enough that p_d is very close to p'. Under these circumstances, the use of

Table 1. Mean square errors (MSE) for selected values of p, m, and k. Underlined values represent minimal MSE, which correspond to optimal k

	Number of composites analyzed (m)						
$p = 0.25$							
k	10	20	30	40	50	70	100
1	0.018750	0.009375	0.006250	0.004687	0.003750	0.002678	0.001875
2	0.012119	0.005720	0.003754	0.002794	0.002225	0.001581	0.001103
3	0.012089	0.004735	0.003039	0.002241	0.001776	0.001255	0.000871
4	0.020238	0.004740	0.002831	0.002055	0.001616	0.001133	0.000782
5	0.042886	0.006821	0.003021	0.002064	0.001597	0.001105	0.000757
6	0.082744	0.015044	0.004519	0.002394	0.001710	0.001144	0.000773
7	0.136580	0.035099	0.010494	0.004130	0.002268	0.001267	0.000826
8	0.197748	0.070379	0.026186	0.010539	0.004834	0.001742	0.000933
9	0.259516	0.119570	0.055926	0.026747	0.013248	0.003914	0.001283
10	0.316952	0.177635	0.100131	0.056860	0.032614	0.011264	0.002904
11	0.367318	0.238401	0.155091	0.101177	0.066229	0.028779	0.008765
12	0.409655	0.296659	0.215018	0.156026	0.113368	0.060140	0.023646
13	0.444165	0.349034	0.274337	0.215729	0.169735	0.105273	0.051727
14	0.471670	0.393946	0.329001	0.274807	0.229593	0.160386	0.093862
15	0.493233	0.431133	0.376772	0.329269	0.287778	0.219896	0.147022
16	0.509932	0.461135	0.416903	0.376893	0.340725	0.278501	0.205891
17	0.522750	0.484879	0.449642	0.416931	0.386589	0.332372	0.265003
18	0.532522	0.503403	0.475770	0.449615	0.424881	0.379403	0.320154
19	0.539937	0.517701	0.496286	0.475717	0.455981	0.418906	0.368851
20	0.545542	0.528649	0.512198	0.496223	0.480727	0.451140	0.410109
$p = 0.10$							
1	0.009000	0.004500	0.003000	0.002250	0.001800	0.001285	0.000900
2	0.005067	0.002450	0.001616	0.001205	0.000961	0.000684	0.000477
3	0.003723	0.001758	0.001152	0.000857	0.000682	0.000484	0.000337
4	0.003087	0.001419	0.000924	0.000685	0.000544	0.000386	0.000268
5	0.002807	0.001224	0.000791	0.000585	0.000464	0.000328	0.000228
6	0.002875	0.001102	0.000707	0.000521	0.000412	0.000291	0.000202
7	0.003493	0.001026	0.000652	0.000478	0.000378	0.000266	0.000184

8	0.005031	0.000986	0.000615	0.000449	0.000354	0.000248	0.000172
9	0.007983	0.000993	0.000591	0.000429	0.000337	0.000236	0.000163
10	0.012906	0.001086	0.000579	0.000417	0.000326	0.000227	0.000156
11	0.020333	0.001360	0.000581	0.000410	0.000319	0.000222	0.000152
12	0.030708	0.001981	0.000610	0.000409	0.000316	0.000219	0.000150
13	0.044328	0.003207	0.000700	0.000416	0.000317	0.000218	0.000148
14	0.061319	0.005380	0.000921	0.000440	0.000321	0.000218	0.000148
15	0.081629	0.008915	0.001400	0.000505	0.000333	0.000221	0.000149
16	0.105051	0.014255	0.002335	0.000660	0.000362	0.000225	0.000151
17	0.131249	0.021832	0.004005	0.000998	0.000433	0.000232	0.000153
18	0.159792	0.032016	0.006759	0.001673	0.000595	0.000247	0.000157
19	0.190198	0.045076	0.010999	0.002910	0.000942	0.000278	0.000162
20	0.221959	0.061156	0.017137	0.005008	0.001623	0.000351	0.000170
$p = 0.05$							
1	0.004750	0.002375	0.001583	0.001187	0.000095	0.000678	0.000475
2	0.002582	0.001253	0.000827	0.000617	0.000492	0.000350	0.000245
3	0.001816	0.000869	0.000571	0.000425	0.000339	0.000241	0.000168
4	0.001428	0.000675	0.000442	0.000329	0.000262	0.000186	0.000129
5	0.001196	0.000559	0.000365	0.000271	0.000215	0.000153	0.000106
6	0.001044	0.000483	0.000314	0.000233	0.000185	0.000131	0.000091
7	0.000941	0.000428	0.000278	0.000206	0.000163	0.000116	0.000080
8	0.000874	0.000388	0.000251	0.000186	0.000147	0.000104	0.000072
9	0.000841	0.000358	0.000231	0.000170	0.000135	0.000095	0.000066
10	0.000850	0.000334	0.000215	0.000158	0.000125	0.000088	0.000061
11	0.000917	0.000315	0.000202	0.000148	0.000117	0.000083	0.000057
12	0.001066	0.000300	0.000191	0.000141	0.000111	0.000078	0.000054
13	0.001332	0.000289	0.000183	0.000134	0.000106	0.000074	0.000051
14	0.001760	0.000280	0.000176	0.000129	0.000101	0.000071	0.000049
15	0.002403	0.000275	0.000170	0.000124	0.000098	0.000069	0.000047
20	0.011151	0.000378	0.000156	0.000111	0.000087	0.000061	0.000041
25	0.035399	0.001608	0.000206	0.000111	0.000084	0.000058	0.000039
30	0.080807	0.007427	0.000801	0.000170	0.000091	0.000058	0.000039
35	0.146936	0.024075	0.004045	0.000752	0.000195	0.000064	0.000040

Table 1. Mean square errors (MSE) for selected values of p, m, and k. Underlined values represent minimal MSE, which correspond to optimal k (continued)

	Number of composites analyzed (m)						
$p = 0.01$							
k	10	20	30	40	50	70	100
1	0.000990	0.000495	0.000330	0.000247	0.000198	0.000141	0.000099
2	0.000524	0.000255	0.000168	0.000125	0.000100	0.000071	0.000050
5	0.000221	0.000105	0.000069	0.000051	0.000041	0.000029	0.000020
10	0.000115	0.000054	0.000035	0.000026	0.000021	0.000015	0.000010
15	0.000080	0.000037	0.000024	0.000018	0.000014	0.000010	0.000007
20	0.000063	0.000029	0.000019	0.000014	0.000011	0.000007	0.000005
25	0.000052	0.000024	0.000015	0.000011	0.000009	0.000006	0.000004
30	0.000047	0.000020	0.000013	0.000009	0.000007	0.000005	0.000003
35	0.000046	0.000018	0.000011	0.000008	0.000006	0.000004	0.000003
40	0.000052	0.000016	0.000010	0.000007	0.000006	0.000004	0.000003
45	0.000074	0.000015	0.000009	0.000007	0.000005	0.000004	0.000002
50	0.000123	0.000014	0.000009	0.000006	0.000005	0.000003	0.000002
60	0.000385	0.000012	0.000008	0.000005	0.000004	0.000003	0.000002
70	0.001088	0.000013	0.000007	0.000005	0.000004	0.000003	0.000002
100	0.010300	0.000118	0.000007	0.000004	0.000003	0.000002	0.000001

equations (2.13) through (2.15) is unnecessary. All of the results presented in this paper consider cases with negligible measurement errors.

In the next sections, we (1) find the optimal value of k, i.e., the number of individual test portions in a composite which minimizes MSE, for a given level of p and a fixed number of analyses; (2) compare the MSE of $\hat{p}_k$ to those of $\hat{p}_1$ for a fixed number of analyses; and (3) compare the costs of achieving a tolerable level of MSE with composite and traditional sampling plans. These comparisons were rapidly accomplished through the use of a computer.

3. Results

The MSE values for combinations of k and m are given in Table 1 for $p = 0.25$, 0.10, 0.05, and 0.01. In general, the MSE decreases to a minimum as the optimal value of k is attained, then increases. The traditional sampling plan for a given number of analyses is represented by $k = 1$. Lower MSE values occur with larger number of analyses (m). After the minimum MSE is reached, the rate of increase in MSE with k decreases as the number of composites increases.

The optimal k (i.e., the number of portions or aliquots in a composite which minimizes the MSE) was calculated for combinations of p and m (Fig. 1). For a given p, the optimal k increases as more analyses are performed. Rapid increases in optimal k are observed when $0 < p < 0.10$. As discussed below, however, previous knowledge of the characteristic of interest

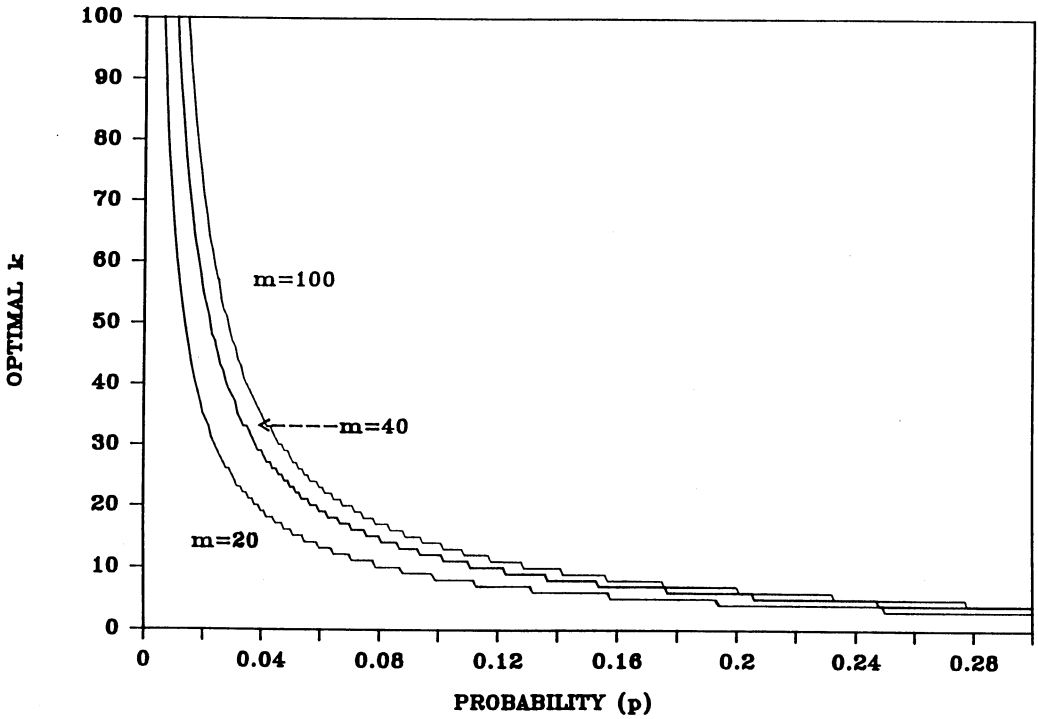

Fig. 1. Optimal composite sample size versus the probability that an individual possesses a characteristic; m equals the number of composite samples analyzed

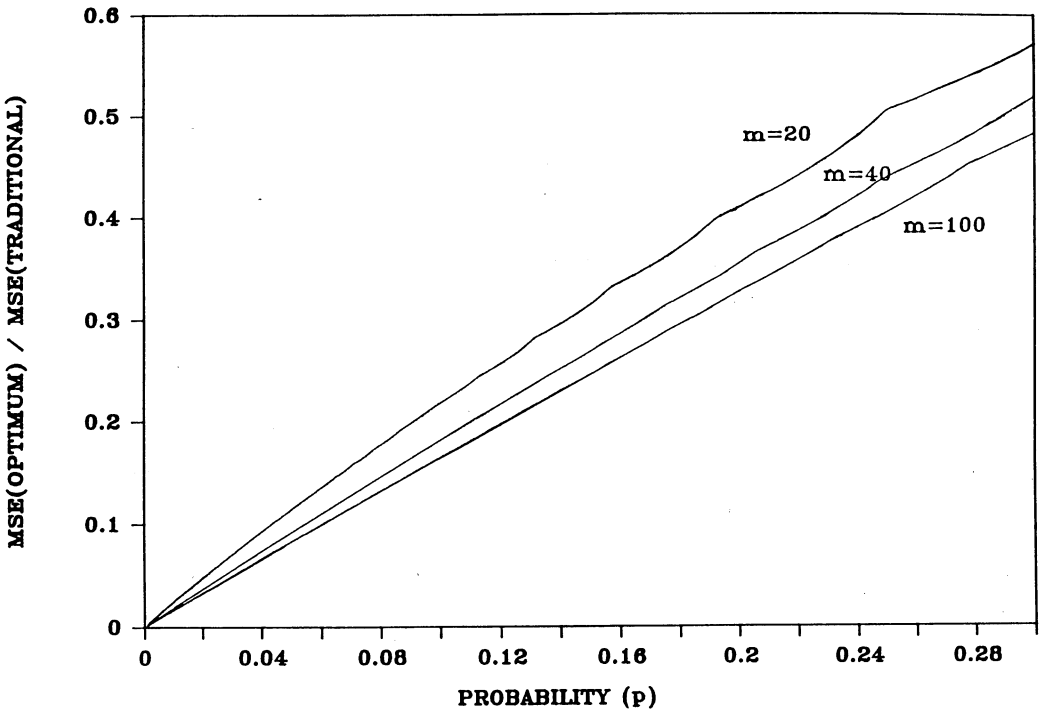

Fig. 2. Ratio of mean square error when optimal k is used to the mean square error when the traditional sampling plan is used versus the probability that an individual possesses a characteristic. Again m equals the number of composite samples analyzed

may place an upper limit on k. As p increases, the family of curves in Fig. 1 converges to an optimal k of 1. Figure 1 is discontinuous because k takes only integer values.

At low values of p, the MSE for optimal k is always less than the MSE for the traditional sampling plan for a given number of analyses (Fig. 2). Clearly, the advantages of sample compositing over traditional techniques are greatest at low levels of p. The discontinuity in Fig. 2 is due to changes in optimal k, which can only assume integer values. In Figures 1 and 2, the distance between the curves for $m = 100$ and $m = 40$ is generally less than the distance between the curves for $m = 20$ and $m = 40$. Increasing the number of composites, m, results in (1) reducing MSE, (2) increasing optimal k and (3) reducing the ratio of optimal MSE to traditional MSE.

4. Discussion

Obviously, the value of p is not known in advance, but typically a researcher can obtain a reasonable upper bound. Establishing such an upper bound before using Table 1 to determine appropriate values of m and k is critical to avoid overestimating optimal k and the large MSE values that result. Clearly, there is also an upper limit to the number of aliquots that can be pooled into a composite. An assumption of the model is that the test portions are homogeneous and representative of the individuals. Furthermore, it is assumed that the aliquots

that make up the composites are representative of the individuals. It is conceivable that below a certain volume, the aliquots may not be representative. If the volume of the composite is fixed, then k is limited by the smallest volume of the aliquot that gives reliable results. The dilution effect has been considered elsewhere (Hwang (1976)). Furthermore, the characteristic of interest may not be detectable when it occurs below a certain concentration or density. For example, if the characteristic of interest is a disease organism that must be grown on a culture medium, there may be a threshold level of population size below which the culture will not have a high probability of becoming established. In such a case, one must be very cautious of large k and small p because the probability of occurrence of the characteristic places a limitation on the number of aliquots that can be pooled into a composite. Previous knowledge of the detectability of the characteristic is necessary.

These techniques can be used to determine the optimal experimental design. Given an experimental goal, such as a maximum tolerable MSE and a preliminary estimate of a reasonable upper bound for p, one can find the value of k which yields the minimum cost. Suppose that in a particular experiment an MSE of 0.001 is the maximum tolerable value when estimating p, and a reasonable upper bound on p is 0.10. Suppose further that the cost of one analysis of test material (C_A), from either an individual test portion or a composite, is four times the cost of obtaining test material from one individual (C_s). The techniques in constructing Table 1 and Figure 1 may be useful in comparing alternative experiments. For example, when $p = 0.10$, and 20 analyses are performed, a design in which $k = 8$ yields an MSE of 0.000986 (Table 1). When 40 samples are analyzed, a design in which $k = 3$ yields an MSE of 0.000857. From Table 1, the traditional plan ($k = 1$) would require approximately 100 analyses to achieve an MSE of approximately 0.001. The costs of these three alternative experiments are estimated as follows:

$$C_k = n(C_s + C_A/k) = mkC_s + mC_A$$

$$C_{k=8} = 20 \cdot 8 \cdot 1 + 20 \cdot 4$$

$$= 240 \text{ cost units}$$

$$C_{k=3} = 40 \cdot 3 \cdot 1 + 40 \cdot 4$$

$$= 280 \text{ cost units}$$

$$C_{k=1} = 100 \cdot 1 \cdot 1 + 100 \cdot 4$$

$$= 500 \text{ cost units.}$$

The first alternative, therefore, is the most cost-effective. It requires only 48% of the cost of the traditional plan to achieve the objective MSE value of 0.001 in this example.

Clearly, the best design for a given experiment depends on p, the tolerable MSE, C_A, and C_s. The researcher can use the techniques described above to select the least costly design that achieves the experimental objectives.

5. References

Cameron, J.M. (1951): The Use of Components of Variance in Preparing Schedules for Sampling Baled Wool. Biometrics 7, pp. 83–96.

Cassie, R.M. (1971): Sampling and Statistics. In Secondary Productivity of Fresh Water, edited by W.T. Edmondson and G.G. Winberg, International Biological Programme Handbook No. 17.

Connolly, J. and O'Connor, R. (1982): Comparison of Random and Composite Sampling Methods for the Estimation of

Fat Content of Bulk Milk Samples. Irish Journal of Agriculture, 20, pp. 35–51.

Gibbs, A.J. and Gower, J. C. (1960): The Use of a Multiple-Transfer Method in Plant Virus Transmission Studies – Some Statistical Points Arising in the Analysis of Results. Annals of Applied Biology, 48, pp. 75–83.

Garner, F.C., Stapanian, M.A., and Williams, L.R. (1987): Composite Sampling for Environmental Motoring. In Principles of Environmental Sampling, edited by L.H. Keith, American Chemical Society, Washington, D.C.

Garrett, R.G. and Sinding-Larsen, R. (1984): Optimal Composite Sample Size Selection, Application in Geochemistry and Remote Sensing. Journal of Geochemical Exploration, 21, pp. 421–435.

Hwang, F.K. (1976): Group Testing With a Dilution Effect. Biometrika, 63, pp. 671–673.

Mack, G. A. and Robinson, P. E. (1985): Use of Composited Samples to Increase the Precision and Probability of Detection of Toxic Chemicals. In Environmental Applications of Chemometrics, edited by J.J. Breen and P.E. Robinson, American Chemical Society, Washington, D.C.

Rohde, C.A. (1976): Composite Sampling. Biometrics, 32, pp. 273–282.

Ryan, J.J., Pilan, J.C., and Leduc, R.J. (1982): Composite Sampling in the Determination of Pyrethrins in Fruit Samples. Journal of the Association of Official Analytical Chemists, 65, pp. 904–908.

Sobel, M. and Elashoff, R.M. (1975): Group Testing With a New Goal, Estimation. Biometrika, 62, pp. 181–193.

Sobel, M. and Groll, P.A. (1959): Group-Testing to Evaluate Efficiently All Defectives in a Binomial Sample. The Bell System Journal, 38, pp. 1179–1252.

Thompson, K.H. (1962): Estimation of the Proportion of Vectors in a Natural Population of Insects. Biometrics, 18, pp. 568–578.

Watson, G.S. (1961): A Study of the Group Screening Method. Technometrics, 3, pp. 371–388.

Received November 1988
Revised December 1989